基于大模型知识蒸馏的专利技术功效词自动抽取方法研究：以车联网 V2X 领域为例

王奎芳 [1,2]，吕璐成 [1,2]，孙文君 [1,2]，王翼虎 [3]，赵亚娟 [1,2]
[1](中国科学院文献情报中心 北京 100190)
[2](中国科学院大学经济与管理学院信息资源管理系 北京 100190)
[3](中国科学技术信息研究所 北京 100038)

摘要：[目的]本文旨在提高专利技术功效自动化提取的准确度。**[方法]**使用 ChatGPT 作为教师模型（Teacher-model），ChatGLM3 作为学生模型(Student-model)，通过知识蒸馏，将 ChatGPT 生成的训练数据微调 ChatGLM3，得到多个技术词抽取模型和功效词抽取模型。采用多个技术词抽取模型分别从专利的摘要、第一权利要求和技术功效语段中抽取技术词，并采用功效词抽取模型从技术功效语段中抽取功效词。**[结果]**微调后的多个技术词抽取模型和功效词抽取模型相较于 ChatGPT，在抽取技术词和功效词时呈现准确率高、召回率低的特点，第一权利要求的 ChatGLM3 微调模型的准确率和 F1 值最高，分别为 0.734 和 0.724。功效词抽取模型抽取的功效词的准确率为 0.649，大于商业工具标注功效词的准确率 0.53。**[局限]**本研究的技术领域和专利语言单一，验证数据量偏小，数据清洗规则还有待于继续优化。**[结论]**本研究方案通过知识蒸馏操作，提升了大语言模型自动化抽取技术功效的准确性。同时，本研究能够支持从专利文本中挖掘前沿创新技术、热点技术，支撑更高质量的智能化专利分析。

关键词： 技术功效词抽取；知识蒸馏；微调大模型；语义相似矩阵
分类号: TP391，G250

Research on automatic extraction of technical and function words extraction method of patent based on large model knowledge distillation: A case study in the field of Vehicle to Everything V2X

Wang Kuifang[1,2]，Lyu Lucheng[1,2]，Sun Wenjun[1,2]，Wang Yihu[3]，Zhao Yajuan[1,2]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China)
[3](Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract：[Objective] This paper aims to improve the accuracy of automatic extraction of key technical words and corresponding function words from patent.**[Methods]** ChatGPT was used as the Teacher-model, and ChatGLM3 was used as the Student-model. Through knowledge distillation method, the training data generated by ChatGPT was used to fine-tune ChatGLM3, and multiple technical word extraction models and a function word extraction model were obtained. The technical words are extracted from the abstract, the first claim and the technical function

paragraph, respectively, by using multiple technical word extraction models, and the function words are extracted from the technical function paragraph by using the function words extraction model.**[Results]** Compared with ChatGPT, the fine-tuned multiple technical word extraction models and function word extraction model show higher accuracy and lower recall rate, when extracting technical words and function words. The ChatGLM3 fine-tuning model of the first claim has the highest accuracy and F1 values of 0.734 and 0.724 respectively. Moreover, The accuracy of the function words extracted by the function word extraction model is 0.649, which is higher than the accuracy of the function words labeled by the commercial tool, which is is 0.53.**[Limitations]** The technical field and patent language of this research are single, the amount of patent verification data is small, and the data cleaning rules expect to be further optimized. **[Conclusions]** This research scheme improves the efficiency accuracy of automatic extraction of large language model through knowledge distillation operation. At the same time, this study can support the mining of cutting-edge innovative and hot technologies from patent texts, and support higher quality intelligent patent analysis.

Keywords: Technical function word extraction; Knowledge distillation; Fine-tuning model; Semantic similarity matrix

1　引言

专利是技术创新成果的重要载体，也是技术情报获取的重要信息来源，全球海量专利数据给技术分析带来巨大挑战。专利作为非结构化文本，在描述方式或术语上并不统一，导致难以使用简单的规则抽取其中的具备创造性的核心技术，当前人工参与标注的抽取方式，已无法满足对大规模专利数据集快速分析的需要。技术功效矩阵或称为技术功能矩阵，是一种典型的专利分析方法，用于发现高价值技术、分析技术热点和空白点、定位特定领域的技术差距等。专利技术功效矩阵对于专利分析很有帮助，但创建起来比较困难。通常，面向中文专利的实体提取方法主要有关键词抽取、实体关系抽取、技术功效主题提取、实体消歧、关键短语、技术主题、知识图谱实体抽取等。就技术功效而言，主要有基于领域知识库或词典[1]、文本分词[2]、基于 TRIZ 理论 [3]、基于 SAO（Subject – Action - Object）结构[4]、句法分析[5]、TF-IDF 算法[6]、基于预训练模型 BART[7]等，以上研究多是借助词性标注或建立词典等半自动化抽取方式。随着大语言模型的发展，ChatGPT 能够理解和学习人类语言并进行对话，推动了人工智能生成技术的应用[8]，利用大语言模型的上下文理解能力，自动生成需要的技术功效内容成为可能[9]。

本文基于大语言模型，通过知识蒸馏操作，采用 ChatGPT 生成的技术词和功效词作为训练语料，对 ChatGLM3 模型进行微调，确定技术词抽取模型和功效词抽取模型，进一步对多种技术词抽取方式进行系统评估，且所抽取技术词与商业工具功效词的准确率比较，确定较为准确的技术词和功效词自动抽取方式。由此提高专利技术功效自动化抽取的准确度，以快速提取专利核心信息，在不需要通读专利全文的情况下，掌握专利的技术和功效要点。实现从专利文本中自动抽取解决特定技术问题的核心技术方案的技术词和功效词，辅助构建技术功效实体，准确反映专利技术发展脉络和分布趋势。

2　相关研究

2.1 文本挖掘技术在技术功效构建的应用现状

随着计算机技术的发展，应用于智能化抽取专利技术功效的技术也在不断更迭，以寻求更加贴近人工标注方式且更加准确的构建方法。2012 年，陈颖等人[10]建立评价单词或短语表示技术功效特征效果的特征度指标，过滤掉数据集中特征度低的词或短语，抽取特征词或技术词。陈晨等人[11]基于人工整合后的德温特专利数据的摘要，使用将文本挖掘与分布式计算相结合的方法，构造技术功效与技术应用矩阵图。2015 年，HE 等人[12]应用语义角色标签创建技术从优势句中提取专利技术和效果短语。翟东升等人[13]使专利摘要构建数据仓库，利用微软数据分析服务，实现技术功效图的构建与多维分析。2016 年，胡菊香等人[14]定位专利摘要中包含技术功效短语的单句，结合依存关系规则、短语规则计算共现频率较高的词，并提取技术功效词。

2017 年，Huang 等人[4]利用斯坦福解析器和关联规则从专利文本的独立权利要求中提取和分离有关技术和功能的信息，构建技术功效矩阵。段庆锋等人[15]研究了基于 SAO（Subject-Action-Object）结构的技术主题、功效主题分析方法，构建摘要的 SAO 技术三元组，抽取技术与功效的词语，经过凝练后构建专利矩阵。2018 年，Amy 等人[16]构建 7 个技术指标和 7 个功能指标，并将挖掘的专利关键术语进行分组。Deng 等人[17]提出一种多特征融合评分算法 PaEffExtr，利用专利效果陈述的分布（效果语句绝大部分出现在摘要的末尾）和形态特征（效果语句中往往有特定的线索词），构造一条线索词库，并使用打分方法从中文专利摘要中自动提取效果语句。

2020 年，王巍洁等人[18]从构成要素、技术工艺与功能效果三个维度，抽取技术词并统计词频，Yang 等人[19]对工艺专利技术词提取，计算候选词的 IF 值和 IDF 值，再从中选择技术词。2021 年，李剑飞等人[20]抽取专利说明书中发明内容部分的技术方案及功效信息，通过相似度计算并辅以阈值筛选建立双方的技术关联关系，最后构建技术-功效图。向姝璇等人[21]从权利要求和说明书发明内容部分抽取核心技术，从说明书背景技术的最后一段、发明内容的第一段或具体说明倒数后几段抽取功效。2022 年，Shi 等人[22]采从中文摘要中综合使用语义依存解析器和预训练的语言模型来提取功能和技术短语。Korobkin 等人[5]通过句法分析的方式从第一权利要求中抽取多个元组，将专利的功能定义为“对象-条件-动作”，并实现非结构化信息的抽取。WANWOOK 等人[23]，提出一种半自动化方式，使用自然语言处理提取专利的关键技术信息，然后将这些信息以矩阵形式可视化形式，该研究仅使用第一个权利要求，因为它通常表达最重要和最详细的信息，并包含总体技术描述，用户可以确认特定专利是否包含所需的技术信息，并可以检测该信息内的关系。

于专利文本信息而言，并非出现频率高就一定是核心技术或功效词，计算词频的方式所抓取的技术词相较于专利真正的核心技术，其准确性有待进一步验证。然而，SAO 结构的抽取分析及主题词的凝练，需要借助专家经验，过程中设置技术指标和功能指标依然离不开人工判断。也有一些研究的侧重点在于对技术功效矩阵做评价指标，将技术和功效两个维度的内容依靠人工解读[24]，将数据归入“技术-功效”矩阵框架中，并没有解决技术功效矩阵构建的耗费大量人力的痛点问题。

从专利文本记载内容来看，专利文本的说明书全文中包括大量信息，如背景技术部分记载技术现状和技术问题的描述，实施方式部分记载技术方案具体内容

展开和扩展描述等，为了避免从专利中抽取信息的杂乱，很少有研究从全文盲目抽取技术或功效。从已公开研究来看，通常专利的技术词主要从摘要或第一权利要求中抽取，功效词主要从摘要、说明书背景技术的最后一段、发明内容的第一段或倒数后几段中抽取。据《专利审查指南》的要求，摘要的字数限制在 300 以内，其通常包括主题名称、第一权利要求的部分内容，有些摘要还包括部分功效内容。而摘要限于字数要求，通常其记载的技术和功效内容上相比于第一权利要求和其说明书中的功效描述都不完整。专利的第一权利要求包括解决技术问题的完整技术方案，发明/实用新型内容的倒数后几段包括对应于第一权利要求的技术功效的较为完整的描述。前述研究涉及从专利的不同内容中抽取技术信息，但是很少用同一个建模方法对专利不同内容中抽取技术词的效果做比较，以在同一标准下推荐最为准确的抽取方式。虽然一些商业化工具，如 Incopat 专利检索数据库，已标注并能导出每条专利的功效词，但其准确性还待进一步验证。专利文本中记载的技术信息量大且语言结构化较差，其中的定义、实体、概念、描述规则等都不统一，有些技术和功效抽取方法难以确定信息的边界、类型等，也为专利技术功效的抽取提出了新的挑战。

2.2　大模型技术在技术功效构建的应用现状

2018 年谷歌团队开创性地提出了预训练语言模型 BERT，之后该模型不断改进，也同时激发了大量的以预训练模型为基础的自然语言处理的应用研究。2023 年，刘春江等人[25]基于 BERT-BiGRU-CRF 抽取技术功能和技术效果的三元组，在不同层级与粒度下自动构建专利技术功效矩阵。2022 年 11 月，ChatGPT 的问世展示了大预言模型的无线潜力，该模型能够理解需求，结合上下文提供合适的答案，也迅速被应用到越来越多的场景中。白如江等人[9]使用 ChatGPT+Prompt 的方法实现专利技术词、功效词以及技术功效二元组的识别、提取和生成。但是，其 prompt 流程中示例的技术词主要来自专利标题内容，技术词的抽取规则比较模糊，虽然每个技术领域检索专利 5000 件，但在每个领域仅仅人工随机标注 50 条数据（其中包括 30 条中文专利、10 条英文专利，10 条日文专利），标注数据量和总数据量差距很大，模型效果有待验证。

中文专利中包含大量的技术信息且描述规则不统一，中文语义多样，借助大模型进行技术词和功效词抽取时难度进一步加大。2023 年 10 月 27 日，中国计算机大会 CNCC2023 上，智谱 AI 发布了自研第三代对话大模型 ChatGLM3[26]，首次加入了代码识别模块 Code Interpreter，在多模态理解、代码生成、网络搜索以及语义和逻辑推理能力都得到了显著增强。由此，本研究结合知识蒸馏（knowledge distillation）方式[27]，以 ChatGPT 作为教师模型，ChatGLM3 作为学生模型，使用 ChatGPT 分别基于专利的摘要，第一权利要求，发明/实用新型内容部分最后几段的技术功效段落生成技术词，并将技术词经过清洗后得到技术词抽取训练数据。那么，从专利的三部分内容中分别得到的训练数据微调 ChatGLM3 模型，得到抽取技术词的微调模型三个，后续经过准确率、召回率和 F1 值对比后，确定准确率最高的微调模型。使用 ChatGPT 从技术功效段落生成的功效词，经过清洗后作为功效词抽取训练数据，对 ChatGLM3 模型进行微调得到抽取功效词的模型。摘要中所记载的功效内容通常包含在技术功效段落中，这里不再对比摘要和技术功效段落在抽取功效词上的效果差异。本文选择将功效词抽取结果与 Incopat 数据库导出功效词进行比较，评估抽取结果的准确率。对

ChatGLM3 微调后的模型进行实证，综合评估抽取效果后确定抽取技术功效更为准确的方式。

3　研究设计

3.1　研究框架

本研究不需要构建领域词典，从人工智能驱动自然语言处理的角度，采用 ChatGPT 和 ChatGLM3 两种大语言模型，以知识蒸馏的方式，构建技术功效抽取方法，提高抽取准确性。本研究的研究框架见下图 1。研究思路主要分为如下三个部分：训练数据处理、模型微调、抽取效果实证。使用 ChatGPT 作为教师模型（Teacher-model），将专利训练数据的第一权利要求、摘要、技术功效句作为输入，为技术词生成定制 prompt，通过 ChatGPT 分别从第一权利要求、摘要、技术功效句中生成技术词，并使用设置的技术词清洗规则对生成的技术词进行优化得到技术词数据 1、技术词数据 2 和技术词数据 3。其中，技术功效句指的是发明/实用新型内容的倒数后几段，对第一权利要求的功效进行描述的段落。为功效词生成定制 prompt，通过 ChatGPT 从技术功效句中生成功效词，并设置功效词清洗规则对生成的功效词进行优化得到功效词数据。将技术词数据 1、技术词数据 2、技术词数据 3 和功效词数据作为训练数据。ChatGLM3 作为学生模型 (Student-model)，通过知识蒸馏操作[27,28]，即将 ChatGPT 生成的训练数据用于 ChatGLM3 模型的学习，基于 P-Tuning v2 微调方法，构建 ChatGLM3 的模型微调。从第一权利要求生成的技术词数据 1，对 ChatGLM3 模型微调后得到技术词抽取模型 1，从摘要生成的技术词数据 2，对 ChatGLM3 模型微调后得到技术词抽取模型 2，从技术功效句生成的技术词数据 3，对 ChatGLM3 模型微调后得到技术词抽取模型 3。基于功效词数据，对 ChatGLM3 模型微调后得到功效词抽取模型。

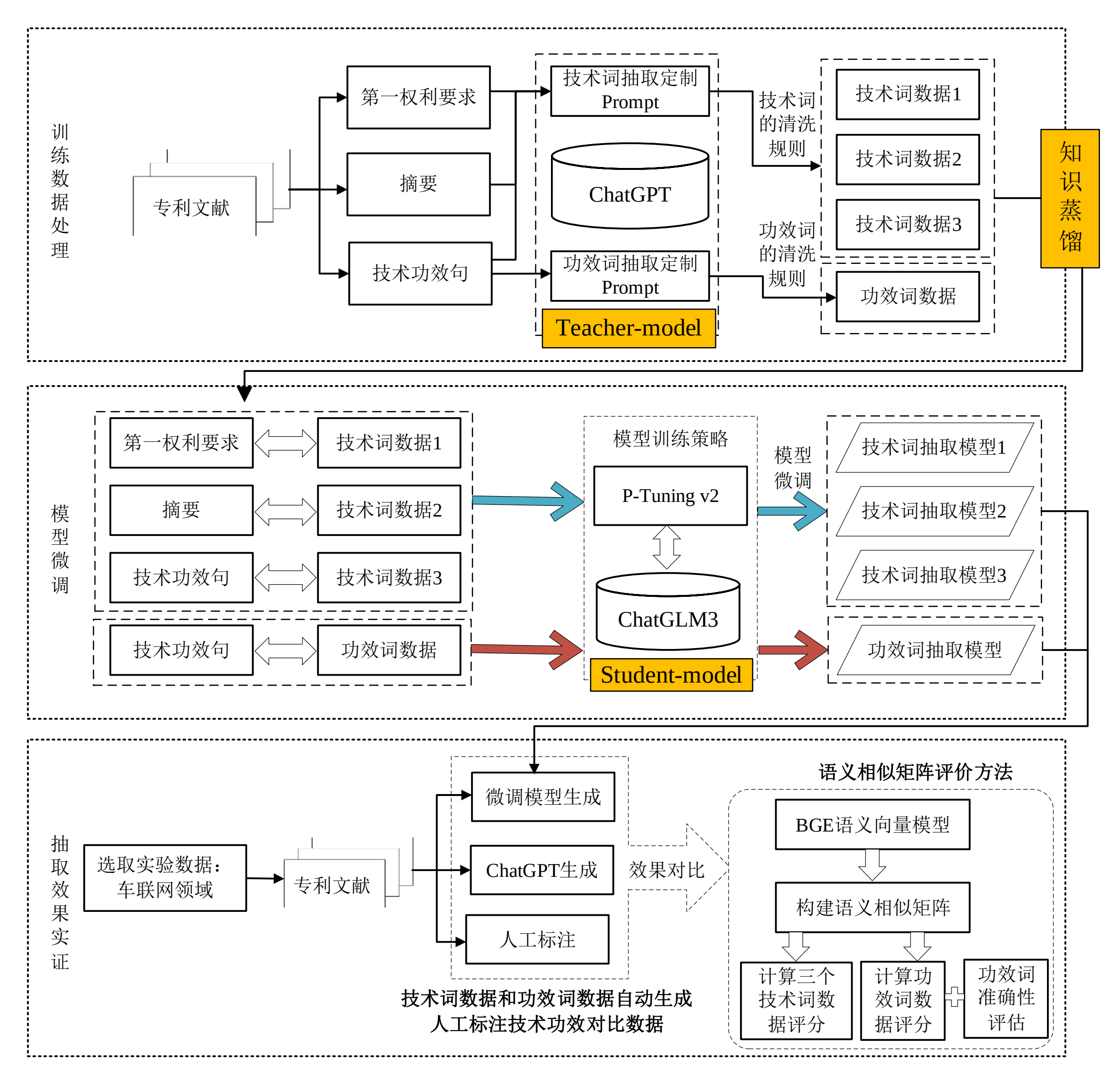


图 1 基于 ChatGPT+ChatGLM3 的技术词和功效词抽取研究框架

Fig1. Research framework of technical and functional word extraction based on ChatGPT+ChatGLM3

使用专利验证数据集，对微调的三个技术词抽取模型和功效词抽取模型的效果进行验证。对技术词而言构建多组技术词数据集，三个技术词抽取模型分别从第一权利要求、摘要和技术功效句中抽取三个技术词数据集，直接使用 ChatGPT 分别从第一权利要求、摘要和技术功效句中抽取三个技术词数据集，人工对每篇专利进行解读标注技术词，以上 7 组技术词数据集基于本研究设计的语义相似矩阵评价方法计算准确率、召回率和 F1 值评分。对于功效词构建多组功效词数据集，分别通过功效词抽取模型、ChatGPT 抽取功效词数据集，人工对每篇专利进行解读标注功效词，以上 3 组功效词数据集基于本研究设计的语义相似矩阵评价方法计算准确率、召回率和 F1 值评分。此外，还将功效词抽取模型抽取的技术词、商业工具（Incopat）导出的技术词与人工标注技术词比较准确率、召回率和 F1 值，以验证有效性。

3.2 数据采集与处理

本文将车联网 V2X（Vehicle to Everything，车用无线通信技术）技术领域的专利作为研究基础，该技术是 3GPP(3rd Generation Partnership Project,第三代合作

伙伴计划)标准组织制定 5G 标准技术系列的重要技术方向，随着智能驾驶技术的发展近年来也受到高度关注。

专利训练数据来自智慧芽全球专利数据库，检索式：(TAC:V2X OR 车联网) AND (DESC:5G) AND (IPC:H04W OR H04L OR H04B OR H04Q OR G08G OR G06F)，合并简单同族，选择包括技术功效语段著录信息的 6278 件中文专利（检索日期：2023 年 10 月）。本研究的训练数据集在同语言同领域下超过现有研究的专利数据量[9]。

实证研究的专利验证数据集来自墨丘标准必要专利（SEP）数据库，用车联网 V2X 技术相关的 15 个技术标准号进行检索，得到 167 件中文专利（2023 年 10 月）。专利验证数据的技术功效句采用从智慧芽专利数据库中导出的技术功效语段或由人工标注。

3.3 技术词抽取方法

（1）ChatGPT+提示 Prompt

ChatGPT 抽取专利训练数据的技术词，技术词从第一权利要求、摘要、技术功效句三种语料中分别抽取。调用 ChatGPT 的 API 进行实验，不修改默认参数。

提示（prompt）相当于一种「提示语」，让 ChatGTP 进入对话模式。根据本文设计的实验框架，设计技术词的提示 prompt。如图 2 所示，技术词抽取任务的 prompt，主要包括：设置信通角色信息，定义技术词的含义，设置输出格式要求，定义输出内容规则。定义技术词是描述专利组件，技术名词的词语或短语。

```
设置系统角色
#        "messages" : [{"role": "system","content": "你是一个多学科专利分析者"},
定义技术词
#                {"role": "system","content": "技术词是描述专利组件,技术名词的词语或短语。下面给定你一段专利文本,请参考上文定义,识别其中的技术词。"},
输出格式要求
#                {"role": "system","content": "以'；'为间隔依次输出每个词语，例如：'车联网能力指示信息；5G毫米波滤波功分模块；V2X连接指示信息；阵列图像传感器；MIMO通信系统；基于传输(TX)资源池的拥塞水平；D2D网络'"},
输出内容规则
#                {"role": "system","content": "要求提取的技术短语长度尽量在3至12个字，请从原文中提取最重要的3-6个技术短语即可，严禁自行概括"},
```

图 2 技术词抽取的 prompt(提示)流程

Fig2. Prompt process for technical word extraction

对 ChatGPT 生成的技术词，通过人工判读总结技术词清洗规则，对技术词进一步清洗。

（2）基于 ChatGLM3+P-tuning 的多个技术词抽取模型

将专利训练数据集的第一权利要求、摘要、功效句三种语料作为 ChatGLM3 模型的输入，将 ChatGPT 预训练得到的技术词数据 1、技术词数据 2、技术词数据 3 作为输出，采用 P-Tuning v2 微调方法，第一权利要求和技术词数据 1 微调 ChatGLM3 得到技术词抽取模型 1，摘要和技术词数据 2 微调 ChatGLM3 得到技术词抽取模型 2，功效句和技术词数据 3 微调 ChatGLM3 得到技术词抽取模型 3。P-Tuning v[29]是深度即时调优的实现，其每个任务有 0.1%到 3%的可训练参数，大大降低了训练时间存储成本和每个任务的存储成本。

使用技术词抽取模型 1,2,3 分别从专利验证数据的第一权利要求、摘要、技术功效句（三种语料）中抽取技术词，得到多组技术词数据集。

3.4 功效词抽取方法

（1）ChatGPT+提示 Prompt

ChatGPT 抽取专利训练数据的功效词，功效词从功效句中抽取。调用 ChatGPT 的 API 进行实验，不修改默认参数。

提示（prompt）相当于一种「提示语」，让 ChatGTP 进入对话模式。根据本文设计的实验框架，设计功效词的提示 prompt。如图 3 所示，功效词抽取任务的 prompt，主要包括：设置信通角色信息，定义功效词的含义，设置输出格式要求，定义输出内容规则。定义功效词是描述专利应用场合，具备的优点，技术所表达功效的词语或短语。

设置系统角色

"messages" : [{"role": "system","content": "**你是一个多学科专利分析者**"},

定义技术词

{"role": "system","content": "**功效词是描述专利应用场合,具备的优点,技术所表达功效的词语或短**语，例如以下示例句：'本申请提供的复合型阻燃母粒在保持较好的雾度和较高的透光率的情况下，不仅具有好的阻燃性能的同时，还具有好的拉伸强度，力学性能优异。'中提取的功效词为：'阻燃性能好；拉伸强度好；力学性能优异'。下面给定你一段专利文本,请参考上文定义,识别其中的功效词。"},

输出格式要求

{"role": "system","content": "**输出格式为：'提高过程可操作性；实现集成化；实现方式有益效果；操作简单；避免遭到损坏'**"},

输出内容规则

{"role": "system","content": "**要求提取的功效词长度尽量在4至10个字，多用动词加名词短语的形式，不要无实义的连词**，以'；'为间隔依次输出每个词语，例如：'提高过程可操作性'、'实现集成化'、'实现方式有益效果'、'操作简单'、'避免遭到损坏'"},

图 3 功效词抽取的 prompt(提示)流程

Fig3. Prompt process for function word extraction

对 ChatGPT 生成的功效词，通过人工判读总结功效词清洗规则，对功效词进一步清洗。

（2）基于 ChatGLM3+P-tuning 的功效词抽取模型

将专利训练数据集的功效句作为 ChatGLM3 模型的输入，将 ChatGPT 预训练得到的功效词作为输出，采用 P-Tuning v2 微调方法，微调 ChatGLM3 得到功效词抽取模型。

使用功效词抽取模型从专利验证数据的技术功效句中抽取功效词。

3.5 专利技术词、功效词抽取效果评估

本研究基于语义相似矩阵的方法，综合评价微调模型的生成效果。由于每个单元的技术词和功效词都包括多个词组，传统对单句进行词重叠计算的评价指标难以较好评价模型效果。本研究采用 BGE（BAAI General Embedding）模型计算各词的语义向量，BGE 是由智源发布的开源中英文语义向量模型，在中英文语义检索精度与整体语义表征能力均超越了社区所有同类模型，同时保持了同等参数量级模型中的最小向量维度，使用成本更低。再计算向量间的 Cosine 余弦相似度，构建语义相似矩阵，余弦相似度的计算公式为 $\frac{x_i^T \hat{x}_j}{\|x_i\|\|\hat{x}_j\|}$，其中 x_i 表示人工标注的文本（技术词或功效词），$\hat{x}_j$ 表示（微调模型或 ChatGPT）抽取的文本，构建的矩阵示意图如图 4 所示，纵轴为人工标注的技术词词组，横轴为模型生成的技术词词组，计算标注词组与生成词组的相似性得出最终分数。

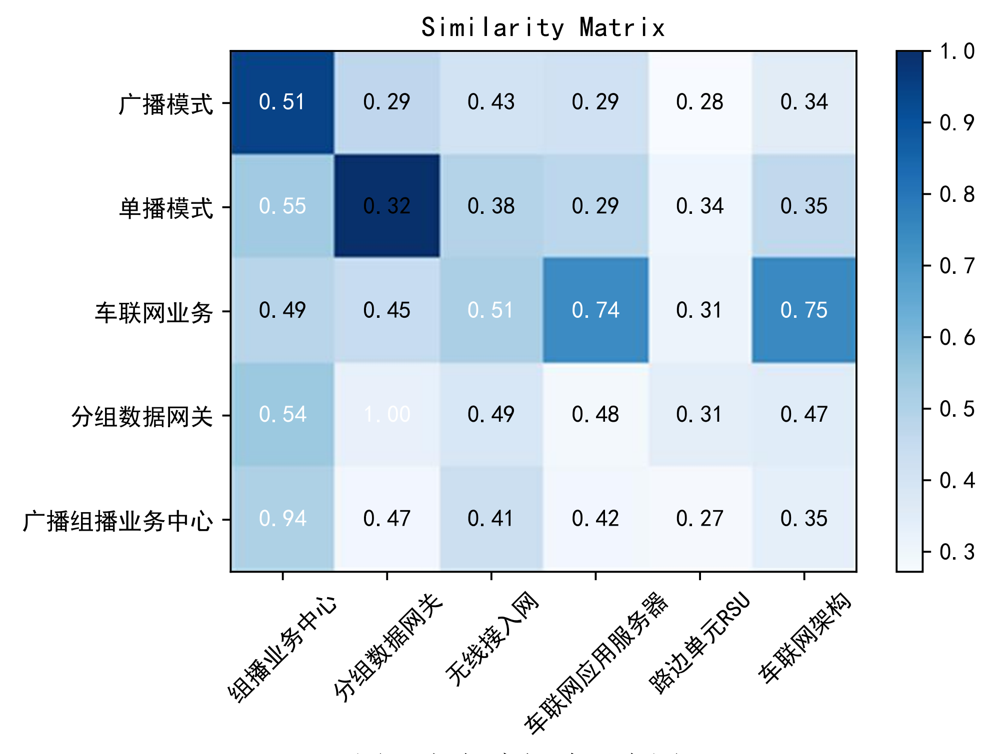


图 4 相似度矩阵示意图

Fig4. Similarity Matrix schematic

相似度结果计算方法参考 BERT Score[30]，准确率、召回率和 F1 值的计算公式如下：

$$R_{BERT}=\frac{1}{|x|}\sum_{x_i\in x}\max_{\hat{x}_j\in\hat{x}}x_i^T\hat{x}_j \quad (1)$$

$$P_{BERT}=\frac{1}{|\hat{x}|}\sum_{\hat{x}_i\in\hat{x}}\max_{x_i\in x}x_i^T\hat{x}_j \quad (2)$$

$$F_{BERT}=\frac{(1+\beta^2)P_{BERT}\cdot R_{BERT}}{P_{BERT}+\beta^2R_{BERT}} \quad (3)$$

在公式（3）中，参数 β 取值越大，整体 F1 值更加关注准确率。由于在专利技术词和功效词抽取的过程中，专利的关键技术推荐追求精确性，尽量不掺杂常规技术手段，抽取结果的正确率往往比抽取数量更为重要，因此在本实验中更加注重准确率，实验过程中 β 取值为 2。

此外，将功效词抽取模型抽取的功效词结果、商业工具 Incopat 导出的功效词结果，分别相对于人工标注的功效词结果计算准确率、召回率和 F1 值。

4 实证研究

将车辆网领域的 6278 件中文专利作为专利训练数据集，用于 ChatGPT 生成技术词和功效词的训练数据，并对 ChatGLM3 进行微调。将选择的 167 件专利作为验证数据集，对微调 ChatGLM3 后得到的三个技术词抽取模型和功效词抽取模型进行验证和效果评估。

4.1 实验环境与参数

本实验环境配置为：CPU，Intel(R) Xeon(R) Gold 6338 CPU @ 2.00GHz；GPU，NVIDIA A100；显存：80GB；Python 版本，3.10.12；Cuda 版本，12.2。实验超参数设置如表 1 所示。选择多个训练步数对比训练效果。

表 1 实验主要超参数设置

Tab1. Experiment main hyperparameter settings

超参数	中文解释	数值
max_source_length	最大输入序列长度	1024
max_target_length	最大输出序列长度	128
train_batch_size	每批次训练数据量大小	1
learning_rate	学习率	2e-2
max_steps	最大训练步数	2000-3000

4.2 多个技术词抽取模型的训练损失对比

训练损失是评估模型在训练数据上的表现的指标之一，其是指模型在每个训练步骤中预测与实际标签之间的差异的平均值，通常希望训练损失随着训练步骤的增加而逐渐减小，这意味着模型在学习更好的表示并更好地匹配标签。对三个从技术词抽取模型（第一权利要求的技术词抽取模型 1、摘要的技术词抽取模型 2、技术功效句的技术词抽取模型 3）的训练损失计算并对比。如图 5 所示，随着训练步数的增加，第一权利要求的技术词抽取模型的训练损失最低。

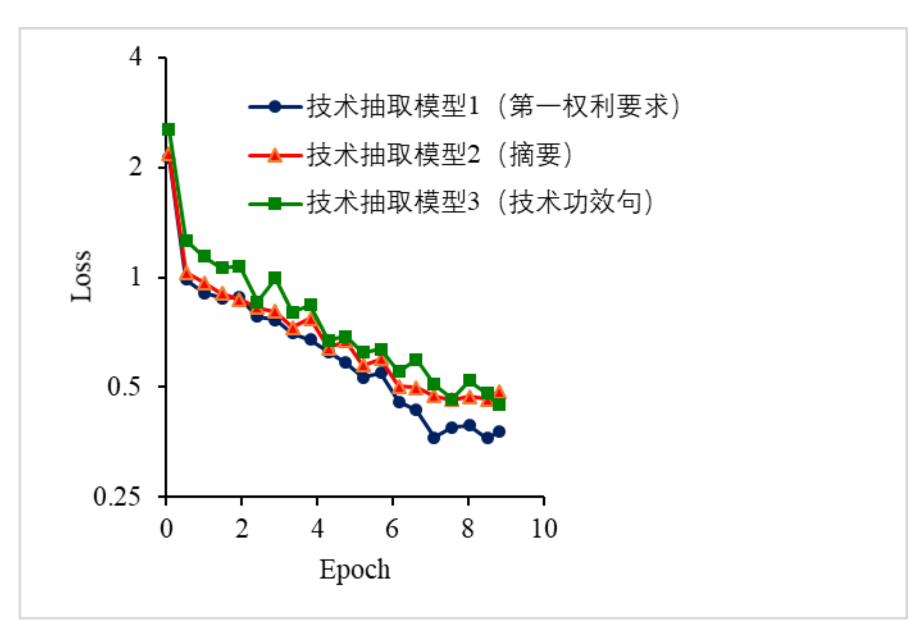


图 5 不同技术词抽取模型训练的损失对比图

Fig5. Loss comparison graph for training of different technical word extraction model

由此，多个技术词抽取模型在训练损失上，第一权利要求表现最佳。

4.3 设置 ChatGPT 生成数据的清洗规则

通过设置清洗规则对抽取的技术词和功效词进行过滤，去除噪声。

（1）技术词清洗规则

本研究在该部分共使用 6278 篇专利，ChatGPT 共生成 45774 个技术词（包括词组或短语），在清洗之前，平均每篇专利生成 7.29 个技术词。清洗掉 16728 个词，最后得到 29046 个技术词，平均每篇专利生成 4.62 个技术词，被清洗掉的为一些噪音词，以提高最终生成技术词的准确性。

如表 2 所示，技术词的抽取规则共 12 个，在表中对每个规则进行了详细介绍和解释，所有规则都在根据 ChatGPT 生成的技术词观察和实验后确定，来删除符合规则限定情况的词语或短语。从表中可以看出，规则 4 清洗掉的词数量最多，为 5287 个，规则 1 和 8 都清洗掉 2000 多个词。例如：“Attention”“预测准确”“网络覆盖率高”“场地数据”等不代表专利技术方案技术手段的无意义噪音词被清洗。

表 2 技术词清洗规则的示例和清洗数据

Tab2. Examples and cleaning data of cleaning rules of technical words

序号	规则介绍	规则解释	示例 1	示例 2	数量
1	只有单个词语	单个词语通常表示某种具体设备或名词，难以概括专利所用技术	Attention	通信端	2288
2	动词+名词+动词	该规则通常表示某种动作的执行，与应用的核心技术有所差别	升级文件下载	调制方式选择	941
3	有形容词且长	该规则由于包含形容词	预测准确	最佳资源	857

	度在 2-5	存在，通常与技术类词语无关			
4	动词+名词且长度在 2-5	该规则主要由动词+名词组合，且长度较短，通常对技术描述不够准确	通信次数	工作时长	5287
5	名词+动词且长度在 2-5	该规则主要由动词+名词组合，且长度较短，通常对技术描述不够准确	任务排序	网络认证	1251
6	数词+名词	该规则由于包含数词，通常对技术描述不够准确	第一主小区	第二设备	1575
7	包含连词的词组	该规则由于连词存在，通常包含两个主体，对核心技术形容不够准确	分析和处理	起点和终点	474
8	包含助词的词组	该规则由于包含助词，与形容词规则类似，对核心技术形容不够准确	语义化的方式	已存储的通信参数值	2493
9	包含时间词的词组	该规则由于包含时间相关词语，对技术的描述不够准确	初始相位	早期测量结果	199
10	只有动词的词组	该规则所有词语均由动词组成，难以准确概括核心技术	开始升级	备案请求	813
11	包含基于、实现、率高、率低等词	该规则通常包含特定非核心技术词语或句式，对核心技术描述不够准确	数据传输速率低	网络覆盖率高	32
12	包含数据一词，且长度在 3-7	该规则通常对数据进行描述，与所用相关技术不符	训练集数据	场地数据	518

（2）功效词清洗规则

本研究在该部分共使用 6278 篇专利，ChatGPT 共生成 34791 个功效词（包括词组或短语），在清洗之前，平均每篇专利生成 5.54 个功效词。清洗掉 6021 个词，最后得到 28770 个功效词，平均每篇专利生成 4.58 个功效词，被清洗掉的为一些噪音词或无意义词，以提高最终生成功效词的准确性。

如表 3 所示，功效词的抽取规则共 6 个，前两个规则属于在句子层面剔除不符合要求的功效词，后四个规则属于在短语层面删除不符合要求的功效词。在表中对每个规则进行了详细介绍和解释，所有规则都在根据 ChatGPT 生成的功效词观察和实验后确定，来删除符合规则限定情况的词语或短语。从表中可以看出，规则 3 清洗掉的词数量最多，为 4053 个，规则 16 清洗掉 1811 个词。例如：“该

段文本中的功效词为：”“识别的功效词为：”“数字编号”“模块化”等不代表专利技术功效的无意义噪音词被清洗。

表 3 功效词清洗规则的示例和清洗数量

Tab3. Examples and cleaning data of cleaning rules o function words

序号	规则介绍	规则解释	示例 1	示例 2	数量
1	去除空值	该规则整篇专利层面去除空值，删除带有 "不能"、"没"、"无" 以及 "功效词" 的短语	示例文段中未提供明确的功效词。	技术问题以及技术效果无法从上述文本中提取出功效词，请提供更多专利文本。	114
2	去除冒号前的文字	该规则去除冒号前的文字	该段文本中的功效词为：	识别的功效词为：	472
3	长度在 4-15	该规则删除长度过短或过长的效果词	云端系统分析实现警示区域的现场警示方案和关联路网警示方案	云控子系统根据秒级导航定位数据	4053
4	限定开头和结尾	该规则删除开头为“实现”、“使用”，结尾为”算法”、“系统”、“策略“、“方法“的词	实现端对端部署策略	实现车联网 ISAC 系统	71
5	细粒度词性组合	该规则删除 n+n 和 v+n 词性组合	发送 SL PRS 的资源	实现 GBR QoS 流的建立	86
6	细粒度词性元素	该规则短语中必须包含动词、形容词或副词中的一种	数字编号	模块化	1811

ChatGPT 生成的技术词和功效词基于上述表 2 和表 3 的清洗规则清洗之后，得到：技术词数据 1、技术词数据 2、技术词数据 3 和功效词数据。

4.4 技术词抽取结果

（1）基于 ChatGLM3+P-tuning 的技术词抽取结果

本文通微调 ChatGLM3 的技术抽取模型能够自动识别和抽取技术词。使用 ChatGLM3 微调得到的三个技术词抽取模型，从专利验证数据集中抽取技术词，示例结果见下表 4。表 4 示出人工标注结果和抽取模型的技术词集合。其中，技术词抽取模型 1 的从第一权利要求中抽取技术词，技术词抽取模型 2 从摘要中抽取技术词，技术词抽取模型 3 从技术功效句中抽取技术词。

表 4 多个技术抽取模型的技术词抽取结果示例

Tab4. Examples of extraction results of technical words of different technical extraction models

ChatGLM3+P-tuning 抽取技术词示例			
人工标注	技术词抽取模型 1	技术词抽取模型 2	技术词抽取模型 3
中央根密钥；第一UE；向网络节点发送请求；识别第一密钥的标识符；私密密钥；直接通信	私密密钥；第二用户设备；第一用户设备；网络节点；直接通信；无线接入网；第一密钥；标识符	根密钥分发；会话密钥获取；UE 邻近	直接通信服务
接收 UE 的身份；分组数据汇聚协议(PDCP)数据单元；用户数据的标识符；PC5 协议；安全链路；发送 PC5 协议消息	安全链路；PC5 协议消息；非互联网协议(非 IP)通信；分组数据汇聚协议(PDCP)数据单元	IP D2D PDCP 封装；非 IP D2D PDCP 封装；设备到设备(D2D)通信；安全数据传输	PC5 信令协议；用户平面传输；PDCP SDU

（2）直接使用 ChatGPT 的技术词抽取结果

直接使用 ChatGPT 分别从专利验证数据的第一权利要求（首权）、摘要和技术功效句中抽取技术词，如下表 5 所示。

表 5 ChatGPT 的技术词抽取结果示例

Tab5. Examples of extraction results of technical words of ChatGPT

ChatGPT 抽取技术词示例			
人工标注	首权抽取	摘要抽取	技术功效句抽取
一个或多个 ProSe 承载；聚合最大比特率(AMBR)参数；建立 ProSe 通信	邻近服务(ProSe)进行通信的方法；无线发射/接收单元(WTRU)；ProSe 承载；聚合最大比特率(AMBR)参数；网络实体；ProSe 通信	邻近服务(ProSe)通信方法和设备；聚合最大比特率参数(AMBR)；服务质量(QoS)；无线发射/接收单元(WTRU)；EPS 承载；分组过滤器；演进型分组系统(EPS)	无线通信系统；无线发射/接收单元(WTRU)；设备到设备(D2D)通信；邻近服务(ProSe)
侧链路单播链路；新的第一下层标识的信息；链路标识符；更新请求消息；新的第一下层标识的侧链路数据帧	侧链路标识；链路标识符；单播链路；下层标识；侧链路数据帧	侧链路单播链路；侧链路标识；下层标识；链路标识符；更新请求消息；更新响应消息；侧链路数据帧	更新用于链路的标识符；单播通信；标识符改变；eV2X 使用；隐私要求；第三方跟踪

4.5 功效词抽取结果

本文通微调 ChatGLM3 的功效抽取模型能够自动识别和抽取功效词。使用 ChatGLM3 微调得到的功效词抽取模型，从专利验证数据集的功效句中抽取功效词。直接使用 ChatGPT 也从专利验证数据集的功效句中抽取功效词。并提供人工解读专利标注的功效词结果。示例数据见下表 6。

表 6 功效词抽取结果示例

Tab6. Examples of extraction results of function words

ChatGLM3+P-tuning 抽取功效词示例		ChatGPT 抽取功效词示例	
人工标注	**模型生成内容**	**人工标注**	**模型生成内容**
支持多个优先级等级；支持多个应用	邻近服务;支持多个优先级等级;支持多种应用	连续性	无线通信系统；设备到设备通信；邻近服务
减少使用的带宽;满足准确性和传输要求	减少带宽使用;提高准确性;满足传输要求	侧行链路通信	提高隐私保护；避免被第三方跟踪；保证服务连续性；防止服务中断

4.6 技术词抽取效果评估分析

（1）微调模型的抽取效果比较

基于 BGE 模型构建语义相似矩阵分析不同模型抽取技术词和功效词的效果。经过微调 ChatGLM3 得到的三个技术词抽取模型分别从第一权利要求、摘要和技术功效句中抽取技术词，直接使用 ChatGPT 也分布从第一权利要求、摘要和技术功效句中抽取技术词，对比两种模型技术词抽取结果（微调模型的训练步数为 3000）和人工标注技术词结果，计算准确率、召回率和 F1 值，结果如表 7 所示。

表 7 技术词抽取效果评估结果

Tab7. Results of evaluation of extraction effect of technical words

模型	训练语料	微调模型抽取结果			ChatGPT 抽取结果		
		Precision	Recall	F1	Precision	Recall	F1
技术词抽取模型 1	第一权利要求	0.734	0.711	0.724	0.703	0.813	0.719
技术词抽取模型 2	摘要	0.662	0.677	0.661	0.645	0.780	0.665
技术词抽取模型 3	技术功效句	0.629	0.593	0.618	0.590	0.653	0.598
平均值		0.675	0.660	0.668	0.646	0.749	0.661

微调 ChatGLM3 得到的三个技术词抽取模型的准确率的计算值，都大于直接用 ChatGPT 抽取技术词得到结果的准确率计算值。除从摘要抽取技术词的微调模型的 F1 值略小于 ChatGPT 之外，其他 F1 值都大于 ChatGPT。三个技术词抽取模型抽取功效词的准确率和 F1 值大于 ChatGPT。微调模型相较于 ChatGPT 呈现准确率高、召回率低的特点。其原因在于 ChatGPT 的抽取词数量较多，涵盖数量较广，因此召回率偏高，但同时噪声数据较多，准确率相对偏低。

专利技术词抽取在三个微调模型上的表现来看，从第一权利要求中抽取技术词的效果最佳，F1 值为 0.724。其原因在于，第一权利要求描述解决技术问题的完整技术方案，涵盖全部必要技术特征，使得微调模型在该语料上的性能最为出色。技术功效句语料更专注于效果提升的描述，而并非技术本身的专业描述，因此更适用于从技术功效句中抽取功效词。摘要语料涵盖了技术背景、技术手段、效果等不同类型的信息，概括更为笼统，且受限于特定字数，也不一定包括完整

技术方案。由此，从第一权利要求中抽取专利的解决技术问题的特点技术特征，表现的效果最好。

（2）超参数对技术词抽取模型 1 的抽取效果影响

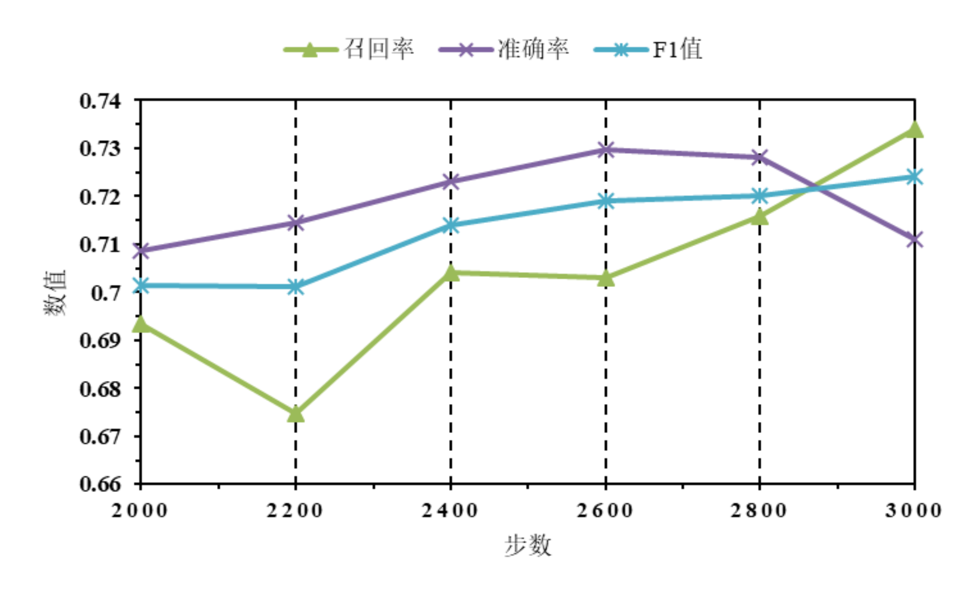


图 6 微调训练步数变化对技术词抽取效果的影响

Fig6. The influence of step number change on technical word extraction

如图 6 所示，改变第一权利要求的技术词抽取模型微调时的训练步数，使用不同步数训练 ChatGLM3 得到的技术词抽取模型分别抽取技术词，探索训练步数对技术词抽取效果的影响。计算不同步数下的技术词抽取模型抽取技术词的准确率、召回率和 F1 值，如图 6 所示，对比发现：2800 步参数下的技术词抽取模型的准确率最大，3000 步参数模型的 F1 值最大。微调 ChatGLM3 时的训练步数并非越大越好，在追求高准确度的情况下，选择 2800 步的训练步数，抽取的技术词准确度较好。

4.7 功效词抽取效果评估分析

如表 8 所示，对微调 ChatGLM3 得到的功效词抽取模型抽取的功效词进行评估，并与 ChatGPT 直接抽取的功效词、Incopat 导出的每篇专利的功效词的评估结果进行对比。计算功效词的准确率、召回率和 F1 值，其中，从准确率结果来看，微调模型的准确率最高，为 0.649，大于 ChatGPT 抽取结果的准确率 0.621，Incopat 标注的功效词仅 0.53。本文基于知识蒸馏的功效词抽取模型的效果优于直接用 ChatGPT 抽取功效词的效果，相比于商业工具的功效词结果，在准确率上有明显提升。此外，功效词抽取的微调模型的召回率、F1 值都高于 ChatGPT 抽取结果和 Incopat 的功效词。其中，Incopat 导出的功效词的 167 件专利的结果中有 35 个空值，去掉空值之后，得到的准确率、召回率和 F1 值分别为 0.602、0.682 和 0.614，也都低于本文微调的功效词抽取模型。由此，通过与大语言模型 ChatGPT 和商业工具对比抽取效果之后证明，本文所微调的功效词抽取模型的效果较好，达到理想预期。

表 8 功效词抽取效果评估结果

Tab8. Results of evaluation of extraction effect of function words

微调模型抽取结果	ChatGPT 抽取结果	Incopat 功效词

Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
0.649	0.792	0.670	0.621	0.776	0.644	0.530	0.592	0.539

通过以上对技术词和功效词抽取效果的系统评估，发现在 ChatGPT 生成语料基础上进行清洗、筛选，可得到更为优质的训练数据，通过知识蒸馏操作，设置 ChatGLM3 模型的微调策略，确定最优训练步数，经微调后的 ChatGLM3 模型，从第一权利要求中抽取技术词具有较高的准确率且效果优于 ChatGPT。

5 总结与展望

本文研究了基于大预言模型知识蒸馏的专利技术功效词自动抽取方法，优化专利技术功效抽取的效果，以提升快速从专利文本识别和抽取技术词和功效词的准确性。系统设计包括训练数据处理、模型微调、抽取效果实证三部分的实验方案，设置知识蒸馏操作，以 ChatGPT 作为教师模型，使用 ChatGPT 从第一权利要求、摘要、技术功效句三种语料中分别抽取的技术词，设计技术词的清洗规则获取优化后更为准确的技术词训练数据。并且使用 ChatGPT 从技术功效句中抽取功效词，并设计功效词的清洗规则获取优化后更为准确的功效词训练数据。再以 ChatGLM3 作为学生模型，使用 ChatGPT 从第一权利要求、摘要、技术功效句抽取的技术词训练数据分别对 ChatGLM3 模型微调，得到三个技术词抽取模型；以及使用 ChatGPT 从技术功效句抽取的功效词训练数据对 ChatGLM3 模型微调，得到功效词抽取模型。最后，使用三个技术词抽取模型和直接使用 ChatGPT 对专利验证数据集的技术词进行抽取，采用 BGE 模型计算抽取结果的语义向量，构建语义相似矩阵，计算抽取结果的准确率、召回率和 F1 值。使用技术功效词抽取模型和直接使用 ChatGPT 对专利验证数据集的功效词进行抽取，获取商业工具标注的专利验证数据集的功效词，同样计算三种功效词结果的准确率、召回率和 F1 值。

抽取结果评估结果表明，三个技术词抽取模型相较于 ChatGPT 呈现准确率高、召回率低的特点，整体表现优于 ChatGPT。从技术词抽取的不同语料角度来看，第一权利要求的技术词抽取模型的训练损失最低，且在模型抽取技术词的效果上，使用技术词抽取模型从第一权利要求语中抽取技术词的效果最佳，而从技术功效句抽取技术词的效果最差。在微调第一权利要求语料的 ChatGLM3 模型时，训练步数选择 2800 时准确率最大，训练步数选择 3000 时 F1 值最大。在功效词的抽取效果方面，本文通过微调 ChatGLM3 得到的功效词抽取模型的准确率、召回率和 F1 值都大于，直接使用 ChatGPT 抽取的功效词以及商业工具标注技术词的准确率、召回率和 F1 值。

本研究方案的通过知识蒸馏微调 ChatGLM3 得到的技术词抽取模型和功效词抽取模型可以优化大语言模型抽取技术和功效的效果，提升抽取结果的准确性。此外，准确的技术词和功效词生成，有助于提供更高质量的专利分析，精准抓取专利的核心技术和效果，快速生成专利技术创新点，掌握技术发展脉络和趋势。

本文的研究内容当前还局限于一个技术领域和一种语言，后续可将本文的模型方法扩展到更多的技术领域和专利语言文本上。在算力允许的情况下，可以进一步扩大专利验证数据集的数据量以将本文的微调模型应用到更多领域的专利文本技术和功效抽取中。此外，还可对生成训练数据的数据清洗规则进一步优化，

比如设计剔除非核心技术词的规则，减少噪声词，优化微调模型，提升抽取结果的效果。

参考文献：

[1]马建红，张明月，赵亚男．面向创新设计的专利知识抽取方法 [J]．计算机应用, 2016, 36(02): 465-471. (Ma Jianhong, Zhang Mingyue, Zhao Yanan. Patent knowledge extraction method for innovation design [J]. Application Research Of Computers, 2016, 36(02): 465-471.)

[2]刘晨．专利信息获取与分析系统关键技术研究 [D]．北京：北京工业大学, 2009.(Liu Chen. Research on Key Technology of Patent Information Acquisition and Analysis System Beijing [D]. Beijing: University of Technology, 2009.)

[3]刘孜．基于多维技术功效图的铂基合金技术机会识别研究 [D]，武汉：华中科技大学, 2019.(Liu Zi. Technology Opportunities Analysis Based on Multidimensional Technology–Function Matrix on Platinum Alloy [D], Wuhan: Huazhong University of Science and Technology, 2019.)

[4] Huang J Y, HSU H T. Technology-function matrix based network analysis of cloud computing [J]. SCIENTOMETRICS, 2017, 113(1): 17-44.

[5] KOROBKIN D M, FOMENKOV S A, KOLESNIKOV S G. A function-based patent analysis for support of technical solutions synthesis [C]. In: proceedings of the 2016 2nd International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), Chelyabinsk, Russia, 2016, DOI: 10.1109/ICIEAM.2016.7911581.

[6] KOROBKIN D M, FOMENKOV S A, KRAVETS A G. Methods for Extracting the Descriptions of Sci-Tech Effects and Morphological Features of Technical Systems from Patents [C]. In: proceedings of the 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA), Zakynthos, Greece. 2018: 1-4.

[7] 邱锐玲．专利关键短语自动提取技术研究 [D]，哈尔滨：哈尔滨工业大学, 2022. (Qiu Ruiling. RESEARCH ON AUTOMATIC EXTRACTION TECHNOLOGY OF PATENT KEY PHRASES [D], Harbin: Harbin Institute of Technology, 2022.)

[8] 钱力，刘熠，张智雄, et al. ChatGPT 的技术基础分析 [J]．数据分析与知识发现, 2023, 7(03): 6-15. (Qian Li, Liu Yi, Zhang Zhixiong et al. An Analysis on the Basic Technologies of ChatGPT [J]. Data Analysis and Knowledge Discovery, 2023, 7(03): 6-15.)

[9] 白如江，陈启明，张玉洁等．基于 ChatGPT+Prompt 的专利技术功效实体自动生成研究 [J]．数据分析与知识发现: 1-15. (Bai Rujiang, Chen Qiming, Zhang Yujie et al. Research on Automatic Entities Generation of Patent Technology Function Matrix based on ChatGPT+Prompt [J]. Data Analysis and Knowledge Discovery: 1-15.)

[10] 陈颖，张晓林．基于特征度和词汇模型的专利技术功效矩阵结构生成研究 [J]．现代图书情报技术, 2012, (02): 53-59. (Chen Ying, Zhang Xiaolin, Research of Patent Technology — effect Matrix Construction Based on Feature Degree and Lexical Model [J]. New Technology of Library and Information Service, 2012, (02): 53-59.)

[11] 陈晨．基于 Mapreduce 计算模型的专利技术—功效—应用图构建与应用研究 [D]．北京：北京工业大学,2013.(Chen Chen. Research on the construction and application of patent technology - Efficacy - Application graph based on Mapreduce calculation model [D]. Beijing: Beijing University of Technology,2013.)

[12] He Y Q, LI Y, Meng L G, et al. A New Method of Creating Patent Technology-Effect Matrix Based on Semantic Role Labeling [C]. 2015 International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI), Beijing, China, 2015: 58-61.

[13] 翟东升,蔡力伟,张杰等.基于专利数据仓库的技术功效图挖掘方法研究——以 3D 打印技术为例 [J]. 现代图书情报技术, 2015, (Z1): 131-138.(Zhai Dongsheng, CAI Liwei, Zhang Jie et al. The Study of Patent Data Warehouse-based Technical Efficiency Map Mining Method -- Taking 3D Printing Technology as an Example [J]. New Technology of Library and Information Service, 2015, (Z1): 131-138.)
[14] 胡菊香,吕学强,刘秀磊等. 专利技术功效短语获取研究 [J]. 科学技术与工程, 2016, 16(14): 228-235. (Hu Juxiang, LV Xue-qiang, Liu Xiu-Lei et al. Extracting Technologies Efficacy Phrases of Patent for Research [J]. Science Technology and Engineering, 2016, 16(14): 228-235.)
[15] 段庆锋,蒋保建.基于 SAO 结构的专利技术功效图构建研究 [J]. 现代情报, 2017, 37(06): 48-54. (Duan Qingfeng, Jiang Baojian. Building Patent Technology － Effect Map Based on SAO Structure [J]. Journal of Modern Information, 2017, 37(06): 48-54.)
[16] TRAPPEY A J C, TRAPPEY C V, GOVINDARAJAN U H, et al. Construction and validation of an ontology-based technology function matrix: Technology mining of cyber physical system patent portfolios [J]. WORLD PATENT INFORMATION, 2018, 55: 19-24.
[17] Deng N, Chen X, Ruan O, et al. PaEffExtr: A Method to Extract Effect Statements Automatically from Patents [C]. In: Proceedings of the 11th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS-2017), Torino, Italy, 2017: 667-676.
[18] 王巍洁, 穆晓敏, 王琰等. 多维专利技术功效分析模型构建及应用研究[J]. 情报理论与实践, 2020, 43(06): 131-134+130. (Wang Weijie, Mu Xiaomin, Wang Yan et al. The Multi-dimensional Patent Technology-effect Analysis Model: Model Construction and Application Study [J]. Information studies: Theory & Application, 2019, 43(06): 131-134+130.)
[19] Yang Y X, Ren G C. Web-based methodology for extracting technology words in Chinese process patents [J]. INTERNATIONAL JOURNAL OF WEB INFORMATION SYSTEMS, 2020, 16(3): 315-329.
[20] 李剑飞,吴红,张彪等. 技术-功效分析视域下的高校专利转移对象识别研究——以石墨烯领域为例 [J]. 情报杂志, 2021, 40(10): 193-199. (Li Jianfei, Wu Hong, Zhang Biao et al. Identification of University Patent Transfer Objects from the Perspective of Technology－Efficacy Analysis——Take Graphene as an Example [J]. Journal of Intelligence, 2021, 40(10): 193-199.)
[21] 向姝璇,李睿. An Improved Technology-Function Features Extraction Method of Patents—An Case Study of 6G Domain [J]. 中国发明与专利, 2021, 18(04): 3-9. (Xiang Shuxuan, Li Rui. Exploration on Automatic Extraction Method of patent technology Efficacy Features: A case study of 6G field [J]. CHINA INVENTION & PATENT, 2021, 18(04): 3-9.)
[22] Zhang C, Mayr P, Lu W, et al. 2022. JCDL2022 workshop: extraction and evaluation of knowledge entities from scientific documents (EEKE2022) [C]. In: Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries, Association for Computing Machinery; Cologne, Germany, 2022: Article 54.
[23] KI WANWOOK, KIM KWANGSOO. Generating Information Relation Matrix Using Semantic Patent Mining for Technology Planning: A Case of Nano-Sensor [J]. IEEE Access, 2017, 5: 26783-26797.
[24] 王学昭, 赵萍, 赵亚娟, et al. “技术-功效”视角下的专利布局形势揭示与风险判定 [J]. 图书情报工作, 2021, 65(16): 73-80. (Wang Xuezhao, ZHAO Ping, ZHAO Yajuan, et al. The Identification of Patent Layout Situation and Risk Based on Technology-Effect Matrix [J]. Library And Information Service, 2021, 65(16): 73-80.)
[25] 刘春江, 李姝影, 刘自强等. 面向多维技术功效分析的专利技术功效矩阵构建方法研究 [J]. 情报理论与实践, 2023, 46(12): 167-174. (Liu Chunjiang, Li Shuying, Liu Ziqiang et al. Research on the Construction Method of Patent Technology/Effect Matrix for Multidimensional Patent Technology/Effect Analysis [J]. Information studies: Theory & Application, 2023, 46(12): 167-174.)

[26] 澎湃新闻·澎湃号·湃客. 国内唯一全面对标 OpenAI 的创业公司，大模型已经出到第三代[EB/OL]. [2023-10-29]. https://www.thepaper.cn/newsDetail_forward_25099097. (Thepaper.cn. OpenAI is the only domestic startup with a comprehensive comparison, and the large model has come out to the third generation [EB/OL]. [2023-10-29]. https://www.thepaper.cn/newsDetail_forward_25099097.)
[27] GOU J P, YU B S, MAYBANK S J, et al. Knowledge Distillation: A Survey [J]. INTERNATIONAL JOURNAL OF COMPUTER VISION, 2021, 129(6): 1789-819.
[28] HSIEH C-Y, LI C-L, YEH C-K, et al. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes [J]. ArXiv, 2023, abs/2305.02301.
[29] LIU X, JI K, FU Y, et al. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks [J]. ArXiv, 2021, abs/2110.07602.
[30] ZHANG T, KISHORE V, WU F, et al. BERTScore: Evaluating Text Generation with BERT [J]. ArXiv, 2019, abs/1904.09675.

致谢：北京墨丘科技有限公司 CEO 黄伟才先生，通过“墨丘科技-全球专利布局分析平台”的标准必要专利数据库，为本研究提供的车联网 5G 声明标准必要专利数据，感谢其在数据上的支持。

通讯作者（Corresponding author）:吕璐成(Lyu Lucheng)，ORCID: 0000-0002-2318-1073, E-mail: lvlc@mail.las.ac.cn。

基金项目:本文系“国家自然科学基金青年科学基金项目”技术距离视角下的技术融合模式、特征及预测研究（项目编号：72304268），基金项目“2023 年度国家资助博士后研究人员计划（C 档）”（项目编号：GZC20232931）和基金项目“支撑科技自立自强的知识产权情报导航分析研究”（项目编号：E329110602）的研究成果之一。
This work is supported by “Research on Technology Convergence mode, Characteristics and Prediction from the Perspective of Technology Distance” under the National Natural Science Foundation Youth Science Fund Project (Grant No. 72304268), “2023 National Funded Postdoctoral Researchers Program (C)” (Grant No. GZC20232931), and the Fund project “Intellectual Property Information Navigation Analysis for Supporting Technological Self-reliance” (Grant No. E329110602).

作者贡献声明：

王奎芳：提出研究思路，设计研究方案，数据采集和分析，撰写论文；
吕璐成：研究思路和研究方案的讨论，论文修订；
孙文君：进行实验计算，结果分析；
王翼虎：进行实验计算，研究方案讨论；
赵亚娟：研究方案讨论。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支持数据：

[1]吕璐成. Technical function data set. DOI: 10.57760/sciencedb.j00133.00404.